García Marín, R.; Alonso Sarría, F.;
Belmonte Serrato, F.; Moreno Muñoz, D. (Eds.) 2016
XV Coloquio Ibérico de Geografía.
Retos y tendencias de la Geografía Ibérica: 1254-1262
Universidad de Murcia-AGE-APG. ISBN: 978-84-944193-4-8

Avaliação e comparação de protocolos de Geocodificação

J. Rocha[1], M.Marques[2], D. Figueiredo[1]

[1] Centro de Estudos Geográficos (CEG) do Instituto de Geografia e Ordenamento do Território (IGOT) da Universidade de Lisboa (UL).
[2] Mapidea, Lda.

Jorge.rocha@campus.ul.pt, miguel.marques@mapidea.pt, danielafigueiredo@campus.ul.pt

RESUMO: O crescente desenvolvimento dos Sistemas de Informação Geográfica (SIG) tem permitido às empresas localizar os seus clientes de modo a adotarem estratégias de Geomarketing espacialmente robustas. Este desenvolvimento dos SIG confere atualmente especial importância às Bases de Dados Geográficas, principalmente no mundo empresarial. Estas bases de dados geográficas são suportadas através das técnicas de geolocalização/ geocodificação.
Normalmente as bases de dados geográficas (TeleAtlas®, Nokia®, Google®, Yahoo!®) utilizadas no geocoding baseiam-se em estruturas topológicas dos segmentos das ruas. Nestas estruturas topológicas podem surgir erros geométricos nos segmentos das ruas, o que pode levar à imprecisão da geocodificação pretendida.
Este trabalho analisa e compara, usando um conjunto de moradas recolhidas aleatoriamente, o comportamento dos diferentes protocolos de geolocalização existentes no mercado. A comparação é feita com referência aos dados georreferenciados manualmente e é analisada tanto em termos quantitativos como qualitativos, tendo como foco Portugal Continental.

Palavras-chave: Geocodificação, Georreferenciação, Erro, Geomarketing.

ABSTRACT: The growing development of Geographic Information Systems (GIS) has allowed companies to locate their clients to adopt geomarketing strategies spatially robust. This development of GIS currently gives special importance to Geographical Databases, especially in the business world. These geographic databases are supported through the geolocation / geocoding techniques.
Typically geographic databases (TeleAtlas®, Nokia®, Google®, Yahoo! ®) used in geocoding based on topological structures of the segments of the streets. In these topological structures can arise geometric errors in the segments of the streets, which can lead to inaccuracy of the desired geocoding.
This paper analyzes and compares, using a set of randomly collected addresses, the behavior of the different geolocation protocols on the market. The comparison is made with reference to manually geo-referenced data and analyzed both in quantitative and qualitative terms, with a focus on Portugal.

Keywords: Geocoding, Georreferencing, Error, Geomarketing.

1. INTRODUÇÃO

O processo de geocodificação é bastante suscetível a erros. Contudo, existe uma abordagem alternativa (ou complementar) de aproximação. Nesta aproximação, são calculados valores estimativos e apresentados em coordenadas geográficas para um determinado local/endereço. Estas coordenadas geográficas são obtidas através do centroíde dos diferentes polígonos (normalmente o centro geométrico), que contêm informação específica e necessária para o processo de geocodificação, nomeadamente a localização do edificado. Associado a estes polígonos estão ainda os códigos postais de 4 dígitos (CP4) ou, ainda mais pormenorizadamente, os códigos postais de 7 dígitos (CP7), que estão estreitamente ligados à identificação codificada da rua, frente de quarteirão e número de polícia.

Ao geocodificar estamos automaticamente de forma simples e dinâmica, a espacializar dados que inicialmente eram meros caracteres. O geocoding é hoje em dia uma das ferramentas utilizada pelas grandes empresas de marketing. Segundo o estudo de Rushton et al., (2008) são várias as empresas que comercializam produtos que contêm informação geográfica, bem como os próprios endereços. São exemplo dessas empresas a Tele Atlas®, a Nokia Here®, a Google® e a Yahoo! ®.

Normalmente as bases de dados geográficas (TeleAtlas®, NOKIA®, Google®, Yahoo!®) utilizadas no geocoding baseiam-se em estruturas topológicas dos segmentos das ruas, nos quais constam os nomes das mesmas. A estas estruturas topológicas estão associados elementos como o número de polícia e código postal correspondentes ao segmento das ruas, normalmente proveniente da identidade dos Correios de Portugal®, ou ainda através dos censos. Nestas estruturas topológicas podem surgir erros geométricos nos segmentos das ruas, o que pode levar à imprecisão da geocodificação pretendida (Rushton et al., 2008).

Segundo Rushton et al. (2008), a geocodificação das localizações requeridas pelo utilizador ao longo de um segmento de recta, são obtidas através da interpolação do número de policia ao longo da linha com o intervalo de números de policia (número de porta) presentes nessa mesma linha. Este processo normalmente funciona bem em meio urbano, mas não tão bem em meio rural, originando falhas constantes aquando da geocodificação. Rushton et al., (2008) verificaram que a maioria das ruas (segmento de recta) na América do Norte, em meio rural apresentavam falhas graves ao nível do intervalo de números de policia (porta), o que levava a falhas constantes no processo de geocodificação.

Por vezes, também é impossível realizar o processo de geocodificação, uma vez que na base de dados geográfica não consta o endereço requerido, ou este apresenta falhas (espaços, ausência de informação como número de porta ou código postal) na própria morada. Em muitos casos estas falhas têm que ser analisadas e corrigidas manualmente pelo utilizador.

Rushton et al. (2008), verificaram que outro dos problemas no processo de geocodificação também passa muitas vezes pela duplicação das moradas presentes nas bases de dados das empresas, originando erros constantes. Através da detecão e posterior resolução dos problemas mencionados no uso da geocodificação, o utilizador irá aumentar os seus conhecimentos na utilização desta ferramenta.

Segundo Goldberg (2008), um dos problemas mais frequentes, é que há bastantes utilizadores que consideram como geocoding determinados métodos que de facto não o são. Muitos investigadores consideram que o processo de utilização do GPS (necessidade de presença física do utilizador no local) para a aquisição de coordenadas geográficas de uma determinada localização, como geocoding. Obter a localização de um determinado objecto através de um sistema de satélites ou através de uma imagen aérea (ortofotomapa) não é geocoding como é muitas vezes afirmado, mas sim georreferenciação. A pesquisa directa de nomes de sítios ou de localizações geográficas em listas por exemplo dos censos, é também erroneamente referida por alguns investigadores como geocoding. O mais comum é falar-se do geocoding como um interpolador baseado em técnicas computacionais que permitem estimar uma localização geográfica através de dados SIG, como por exemplo os ficheiros vectoriais de uma rede viária ou ainda parcelas constituídas por vários ficheiros vectoriais (Goldberg, 2008).

Outro dos problemas encontrados por Goldberg (2008) diz respeito ao facto do geocoding poder ser encontrado em diversos softwares SIG. Para o autor, o processo de geocoding pode ser visto como uma operação única, no entanto esta necessita de diversos algoritmos, operações e diferentes fontes de dados que vão "trabalhar" mutuamente para chegar a um output final. Cada componente inserido é resultado de uma pesquisa feita em várias áreas científicas. Assim, coloca-se a questão, se realmente quando se fala na expressão geocoding, nos referimos ao processo como um todo, ou apenas a algum (uns) dos componentes que o constituem.

Portanto, parece óbvio que a utilização do processo de geocodificação leva muitas vezes a erros e a alguma falta de precisão na obtenção das localizações desejadas. Assim, Goldberg (2008) menciona que após a fase de geocodificação é essencial haver uma revisão manual por parte do utilizador, de modo a validar os resultados obtidos. No entanto, este processo tem como desvantagem o tempo que é necessário dispender para realizar essa tarefa. Os erros mais comuns são os na entrada dos dados, principalmente nos endereços das respectivas localizações, sendo estes de fácil correcção manual. Estes erros são um "quebra-cabeças" para os softwares e fáceis de descortinar para o utilizador.

2. METODOLOGIA

L A geocodificação de 2140 moradas escolhidas aleatóriamente, foi efectuada através do Google Earth tendo sido posteriormente validado através da confrontação com as coordenadas provenientes da base Nokia

Here.

De modo a validar os resultados obtidos foi necessário ter em conta que os pontos provenientes da Nokia tinham diferentes níveis de precisão (Figura 1), nomeadamente:

Nível de Precisão 1 – Corresponde à geocodificação efetuada ao nível do número de polícia;

Nível de Precisão 2 – Corresponde à geocodificação efetuada ao nível do segmento da rua;

Nível de Precisão 3 – Corresponde à geocodificação efetuada ao nível do centro da rua;

Nível de Precisão 4 – Corresponde à geocodificação efetuada ao nível do centro da localidade;

Nível de Precisão 5 – Corresponde à geocodificação efetuada ao nível do centro do Município/freguesia.

Assim a validação efetuada teve em conta a confrontação entre os pontos obtidos pelo processo de geocodificação (Google Earth) e os pontos por diferentes níveis de precisão provenientes da Nokia.

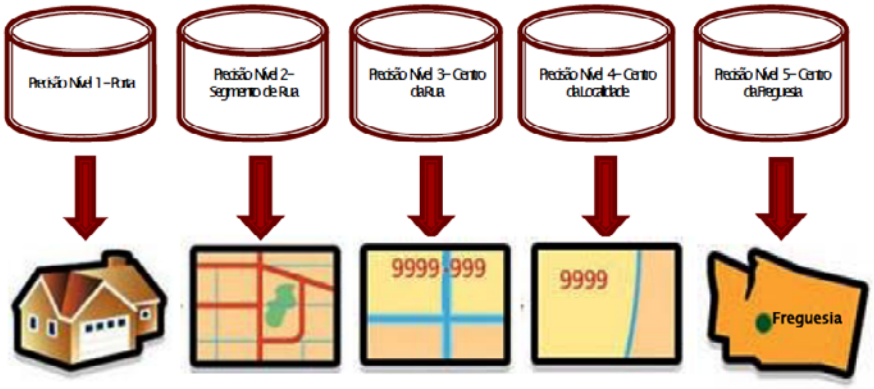

Figura 1. Diferentes níveis de precisão como resultado do processo de geocodificação (manual ArcGIS online, ESRI).

Actualmente a maioria das empresas armazena informação dos seus clientes (nome, endereço, código postal, longitude /latitude) em tabelas formatadas para o efeito. Assim através do geocoding a empresa vai conseguir ter automaticamente e quase instantaneamente a localização (x,y) – dos dados previamente tratados (normas de reconhecimento). Estes mesmos dados poderão ser espacializados num mapa digital.

Tendo esta informação adquirida é possível às empresas levar a cabo melhores estratégias de marketing com "targets" específicos, de modo a definir-se rumos eficazes de actuação.

Na prática o geocoding não é mais que uma tradução de uma descrição em texto (e.g. morada) de uma localização, para valores de coordenadas, permitindo ter pontos no mapa.

É através dos localizadores de endereços (adress locators), que são construídos através de normas definidas pelo utilizador (adress locator styles), que irá ser dada consistência (ou não), à transformação dos endereços em coordenadas geográficas. Quanto maior for a qualidade dos dados de input, melhor serão os resultados finais (Figura 2).

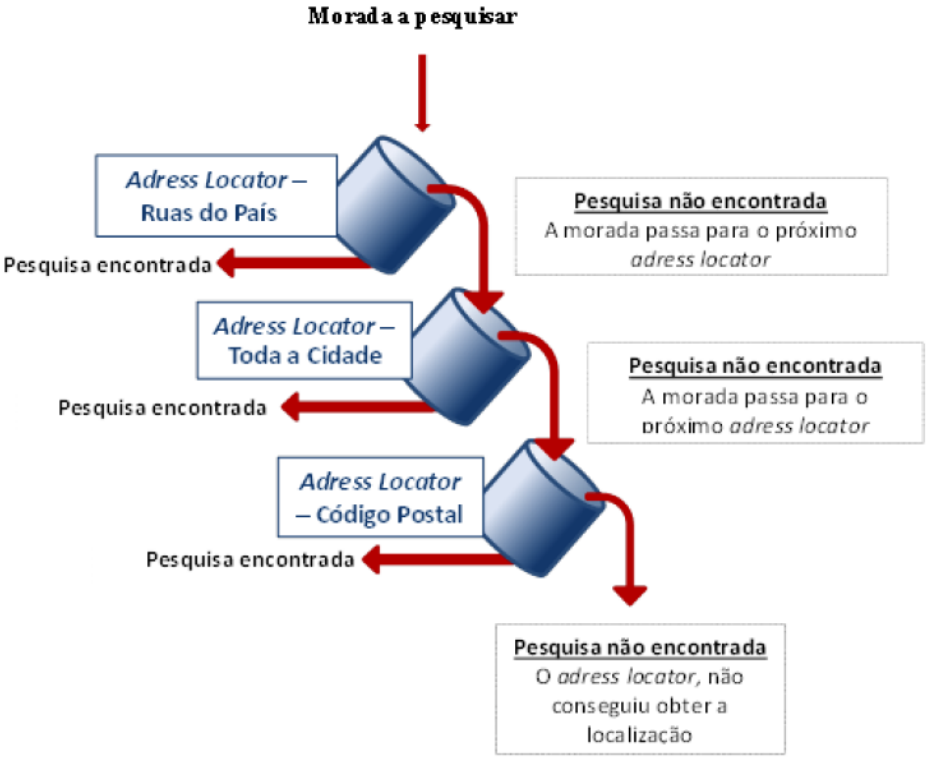

Figura 2. Processo de formação de um adress locator (manual ArcGIS online, ESRI).

3. ANÁLISE DE RESULTADOS

O nível de precisão que apresenta menor percentagem de erro é o nível de precisão 1 (12%), que corresponde à geocodificação efetuada ao nível do número de polícia.

Relativamente ao nível de precisão 2, assiste-se a um aumento significativo do erro, passando dos 12% do nível 1 para 45%. Os resultados obtidos demonstram que a partir do nível 2 a fiabilidade da Nokia começa a decair. Esta afirmação é reforçada ainda pelo facto de nos níveis seguintes observar-se percentagens cada vez maiores, nomeadamente no nível de precisão 3 (70%), no nível de precisão 4 (100%) e por fim o nível de precisão 5 também com uns expressivos 100%.

Estes resultados podem ainda ser analisados através da medida de Regressão Linear. Recorrendo à mesma é possível medir a força e direção num relacionamento entre duas variáveis, ou seja, a dependência que a variável "X" apresenta em relação à variável "Y" (Ebdon, 1982). A análise aos resultados obtidos demonstra que existe uma forte relação positiva entre a variável "Percentagem de Erro" e os "Níveis de Precisão do AVGR".

Por outro lado, através do rácio (ratio) entre as duas variáveis irá ser possível medir a qualidade da dependência entre as variáveis tidas em consideração no estudo. O cálculo deste ratio é feito através da seguinte equação (Ebdon, 1982):

$$r^2 = s_{\hat{y}}^2 / s_y^2 \quad (1)$$

Em que r^2 é o coeficiente de determinação; $s_{\hat{y}}^2$ é a variação explicada de Y e s_y^2 é a variação total de Y. O coeficiente de determinação (r^2) varia entre 0 e 1. Quanto mais próximo da unidade estiver o Coeficiente de Determinação, maior será a validade da regressão.

Relativamente à validação da informação geolocalizada, o valor do Coeficiente de Determinação (r^2) foi de 0,9381 (Figura 3), ou seja à volta de 94%. Sendo o valor do Coeficiente de Determinação próximo de 1, significa que há uma forte dependência entre as variáveis analisadas, i.e. à medida que o nível de precisão é menos exigente (de 1 para 5) a percentagem de erro aumenta.

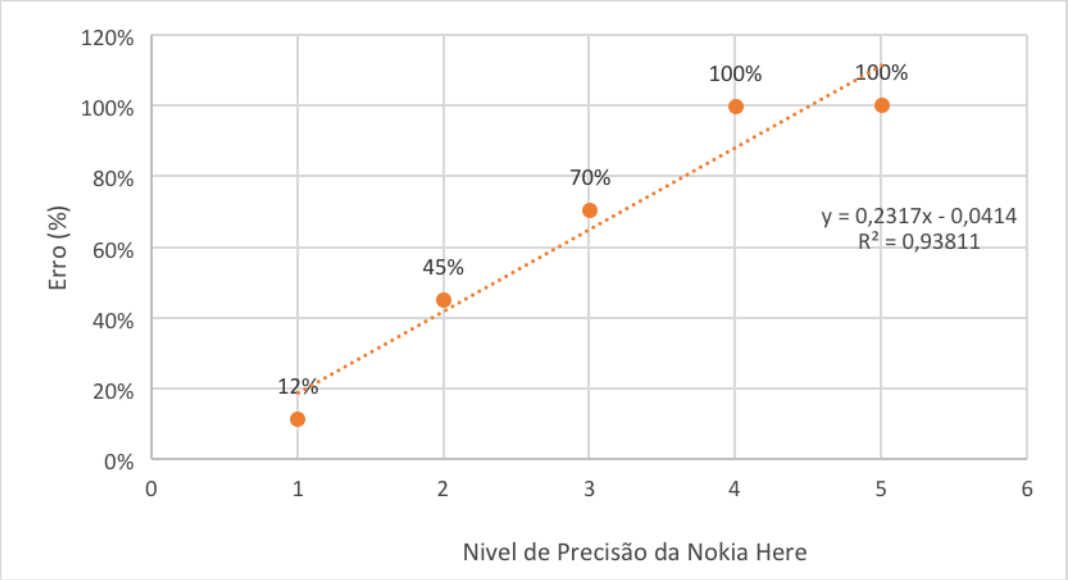

Figura 3. Reta de regressão do Erro de Geolocalização.

Através da análise dos resultados obtidos por nível de precisão (Figura 4), é possível descortinar qual o nível de precisão que é mais fiável utilizar no ato da geocodificação, tendo como referência a base da Nokia (geocodificação).

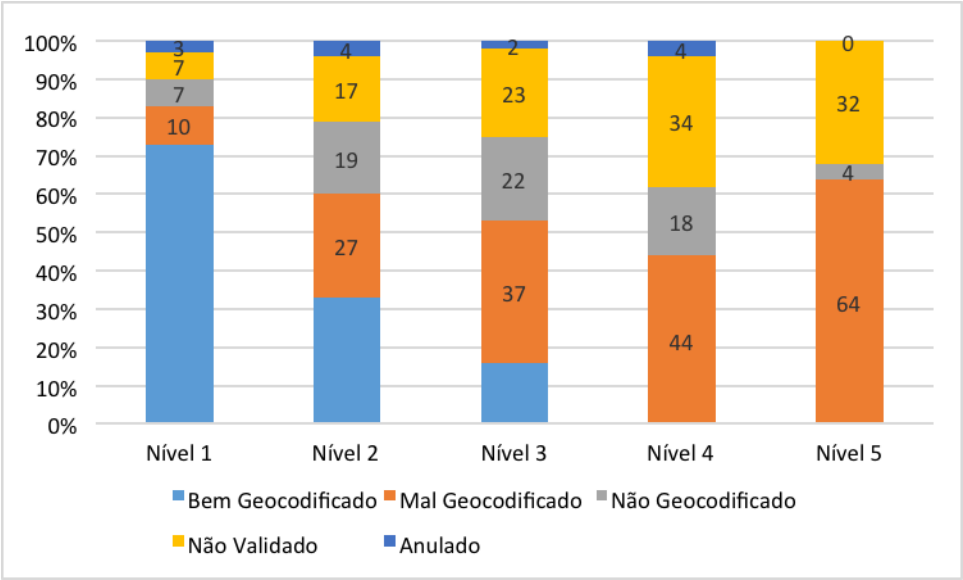

Figura 4. Desemprenho da Nokia Here segundo o nível de precisão.

Da validação entre o processo de geocodificação (Nokia), verificou-se que os níveis de precisão 1 e 2 eram os que apresentavam maior fiabilidade, tendo sido estas as localizações que foram tidas em consideração para a análise no GooGle. Os resultados obtidos constam da Tabela 1.

Tabela 1. Erros apresentados pelo Google

Localizações	*Precisão*	*Precisão*

(Google)	*Nível 1*	*Nível 2*
Bem Geocodificado	542	230
Mal Geocodificado	86	102
Total	628	332
Erro (%)	14	31
Sem validação (not found)	6	5

É de referir ainda que neste processo de geocodificação foram feitas várias simulações de modo a descobrir de que forma o Google permitia um maior número de geocodificações e menos "not found".

Tendo em consideração os resultados obtidos podemos verificar que tanto no Nível de precisão 1 como no Nível de precisão 2, o número de localizações "Bem Geocodificadas" é superior às localizações "Mal Geocodificadas", obtendo percentagens de erro de 14% e 31% respetivamente.

Quanto às localizações não encontradas também apresentam um baixo número de ocorrências, 6 (precisão de nível 1) e 5 (precisão de nível 2). No que diz respeito à percentagem de erro por nível de precisão, o nível de precisão 1 apresenta uma taxa bastante considerável de 85% localizações "Bem Geocodificadas", tendo 14% de localizações "Mal Geocodificadas. Por sua vez o nível de precisão 2 conta com 69% de localizações "Bem Geocodificadas" e 31% de "Mal Geocodificadas".

É de notar ainda que os erros mínimos tanto num software como no outro, verificam-se nas áreas metropolitanas de Lisboa e Porto, o que mostra o bom detalhe oferecido por parte dos fornecedores de BD geográficas para as principais ciudades portuguesas.

Através do cruzamento da espacialização dos dois mapas relativos aos dois processos de geocodificação (AVGR e MMQGIS), foi possível perceber quais as áreas em que o software do AVGR tem maior precisão em relação à base de dados geográfica do Google® (MMQGIS), e vice-versa (Figura 5). Este cruzamento foi feito através da ferramenta do ArcGIS raster calculator, onde se subtraiu o output da NOKIA (AVGR) (valores das distancias das localizações mal geocodificadas) do output com os valores das distâncias das localizações mal geocodificadas do MMQGIS.

Analisando a Figura 5, verifica-se que o AVGR é relativamente melhor nas zonas de Vila Real, Bragança, Coimbra, em certas partes de Santarém, Lisboa e Setúbal, bem como em Faro. Nas restantes áreas predomina a BD geográfica do Google® (MMQGIS). Estes resultados mostram a boa fiabilidade que a base de dados geográfica do Google® (MMQGIS) tem, pois mesmo nas zonas assinaladas a vermelho (AVGR melhor), o erro máximo nessas áreas é algo relativo (e.g. Bragança – 16km).

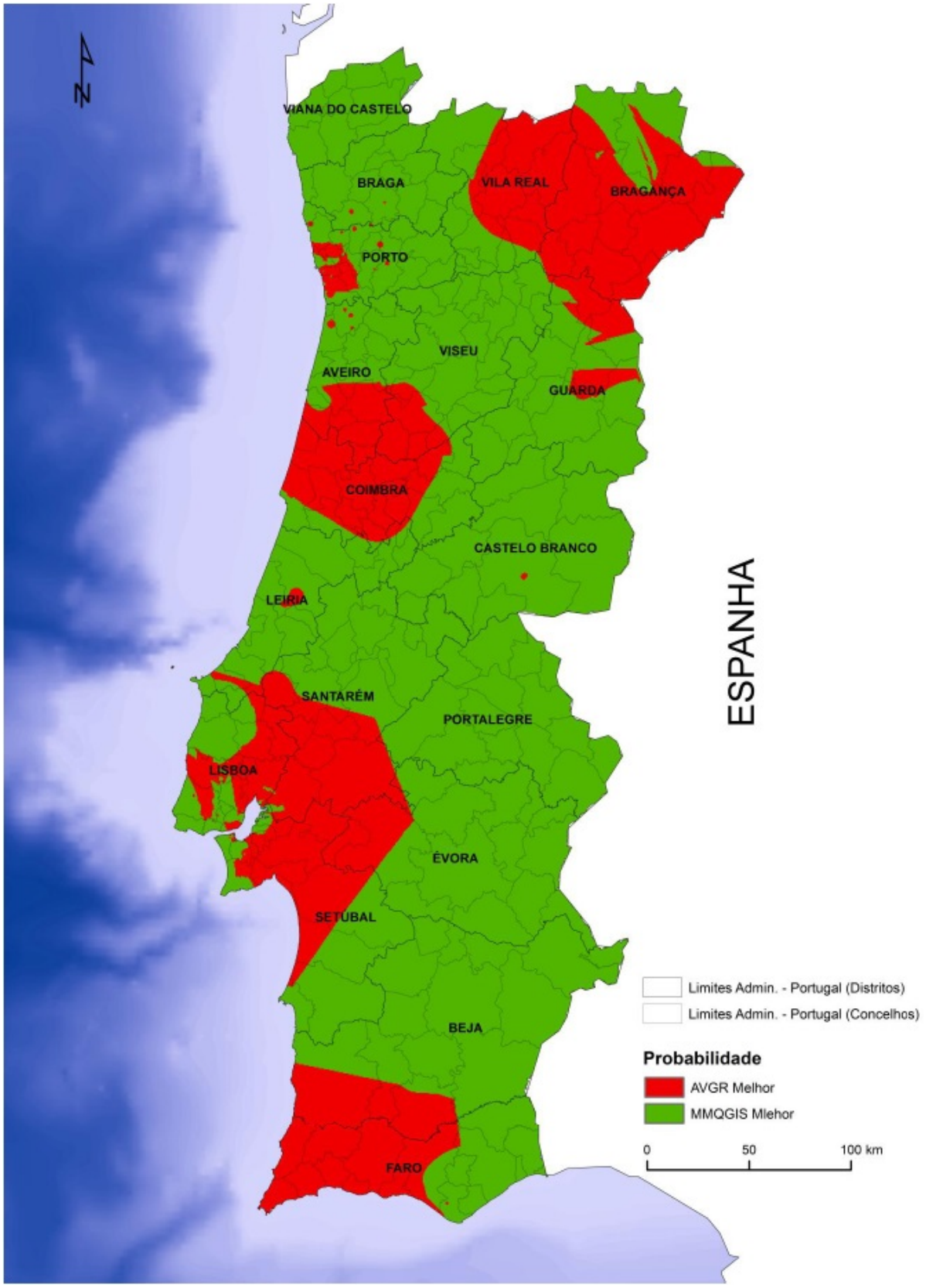

Figura 5. Cruzamento entre os resultados do MMQGIS versus AVGR.

4. DISCUSSÃO

Comparativamente à validação que foi feita entre o processo de geocodificação efetuado pela Nokia, onde foram distinguidas as "Localizações (não) próximas do Ponto Estimado", a base da Goolge conseguiu geocodificar mais 45 localizações no nível de precisão 1 e mais 74 localizações no nível de precisão 2.

Para além do processo de geocodificação foi possível confrontar também os resultados obtidos pela Nokia e pelo Google em relação aos erros máximos, mínimos e médios. Os resultados deste processo podem ser observados na Figura 6.

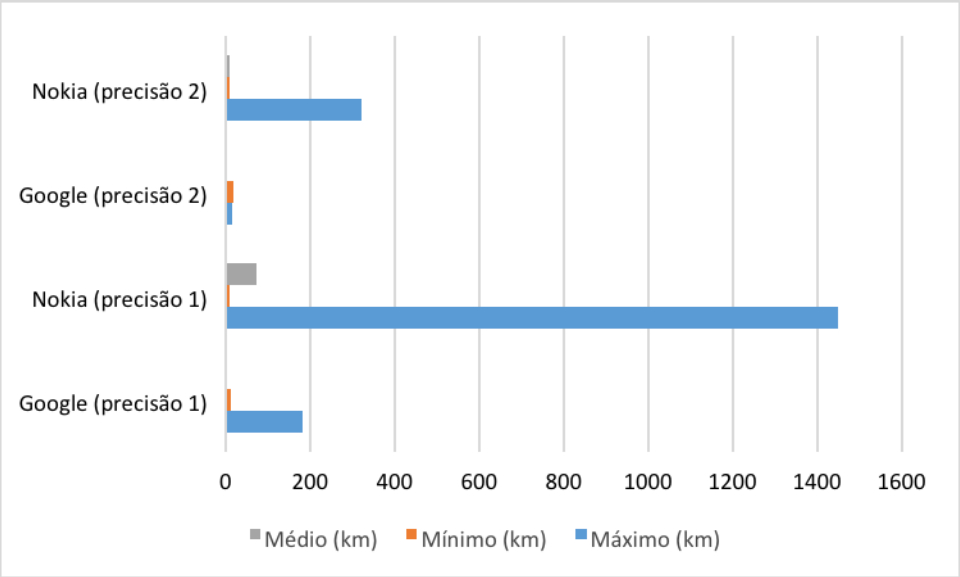

Figura 6. Comparação dos erros entre as duas plataformas.

Através dos resultados obtidos relativamente à distância do erro máximo, mínimo e médio, verifica-se a vantagem que o Google, tem em relação à Nokia. Verifica-se apenas uma exceção no que diz respeito ao erro mínimo, onde apesar dos desempenhos serem favoráveis à Nokia, os valores estão muito próximos dos da base geográfica do Google®.

É de notar ainda que os erros mínimos tanto numa plataforma como noutra, ocorrem nas áreas metropolitanas de Lisboa e Porto, o que mostra o bom detalhe oferecido por parte dos fornecedores de BD geográficas para as principais cidades portuguesas.

Através do cruzamento da espacialização dos dois mapas relativos aos dois processos de geocodificação (Nokia e Google), foi possível perceber quais as áreas em que a Nokia tem maior precisão em relação à base de dados geográfica do Google, e vice-versa. Este cruzamento foi feito através da ferramenta raster calculator do ArcGIS 10.2, onde se subtraiu o output da Nokia (valores das distancias das localizações mal georreferenciadas) do output com os valores das distâncias das localizações mal georrefenciadas do Google.

Verifica-se que a Nokia é relativamente melhor nas zonas de Vila Real, Bragança, Coimbra, em certas partes de Santarém, Lisboa e Setúbal, bem como em Faro. Nas restantes áreas predomina a BD geográfica do Google. Estes resultados mostram a boa fiabilidade que a base de dados geográfica do Google®tem, pois mesmo nas zonas assinaladas a Nokia apresenta melhor desempenho, o erro máximo é algo relativo (e.g. Bragança – 16km).

5. CONCLUSÕES

O sucesso da geocodificação está estreitamente ligado à capacidade de match rate que este processo pode ter. Significa isto que quanto maior for a percentagem de localizações bem geocodificadas, melhor será a precisão de analise por exemplo da disseminação de uma doença. Outro dos factores também não menos importante no processo de geocodificação é a precisão de posicionamento das localizações geocodificadas. Isto significa que é sempre necessário observar se a localização estimada pelo processo de geocodificação tem um posicionamento perto ou afastado da realidade.

O software (AVGR) permite proceder ao processo de geocodificação. Este software possui como base de dados Geográfica a NOKIA® (serviço pago). Assim deveria-se apostar em software que tivesse como suporte a base de dados geográfica da Google®, pois esta possui uma menor percentagem de erro aquando do processo de geocodificação. No entanto é de referir que apesar do utilizador poder usar a base de dados geográfica da Google® de maneira livre, esta tem o limite de consultas de 2 500 solicitações de geolocalização por dia. Caso o utilizador, e neste caso mais direccionado para o mundo empresarial, pretenda realizar um maior número de consultas por dia, existe a possibilidade de aquisição de uma licença da Google® de modo a serem. possíveis 100 000 consultas por dia. O limite de consultas é imposto pela Google®, de modo a não haver abusos, nem adulterações no código da API.

Só através da conjugação de métodos o utilizador irá conseguir ter uma boa fiabilidade nos seus resultados.

6. BIBLIOGRAFÍA

Ebdon D. (1985): Statistics in Geography: Second Edition. Cambridge, USA

Goldberg, D. (2008): A Geocoding. Best Practices Guide. GIS Research Laboratory. University of Southern California

Rushton, G.; Armstrong, M.; Gittler, J.; Greene, B.; Pavlik, C.; West, M.; Zimmerman, D. (2008): Geocoding Health Data: The use of Geographic Codes in Cancer Prevention and Control, Research, and Practice. CRC Press